Supplementary material to the article: Estimating the structural segmentation of popular music pieces under regularity constraints

Gabriel Sargent, Frédéric Bimbot, Emmanuel Vincent

Supplementary material to the article: Estimating the structural segmentation of popular music pieces under regularity constraints

Gabriel Sargent, Frédéric Bimbot and Emmanuel Vincent

I. Short description of the MIREX structural segmentation systems used in the article

This section gives a short description of the structural segmentation systems referenced in Section V (Table III) and considered for fusion in Section VII.C. They constitute the most representative systems submitted to MIREX between 2010 and 2015, i.e., 17 over 40 submissions (including IRISA10-11-12). They were selected according to their performances and/or their specificities. Also, as some systems were submitted several times with some variations in their parameter values, we ignore the duplicates that obtained similar or lower performances for border estimation. We provide below a short description of the structural boundary estimation process of each system[1] and sorted according to their assumptions on the type of structural segments.

Systems CC1, CL1, GP6-7, KSP1 and RBH2 rely essentially on homogeneity criteria :

- System CC1 by Cannam *et al.* implements the approach of [1]. It represents the timbral content of a music piece by a sequence of beat-synchronous histograms, each histogram describing the distribution of low-level timbral states around each beat. The structural segmentation is performed by a soft K-means algorithm adapted to favor the grouping of temporally close histograms [2], K being fixed beforehand to a small value [1].
- System CL1 by Chen and Li first produces multiple homogeneity-based structures from a sequence of MFCCs and a sequence of Chroma vectors using a multi-level clustering incorporating a turbulence module. A score matrix merging all these structures is computed then segmented using Non-negative Matrix Factorization to obtain the estimation of the structural boundaries [3].
- Systems GP6 and GP7 by Peeters and Cornu both estimate the structural borders through the analysis of the weighted sum of similarity matrices calculated from three sequences of timbral features vectors (MFCCs and spectral moments) and a sequence of tonal features (Chroma vectors). The homogeneous zones of the resulting matrix are then searched to segment the corresponding music piece [4].
- System KSP1 by Kaiser *et al.* describes a music piece by a sequence of MFCCs and a sequence of tonal features emphasizing the transitions between the pitch classes of adjacent Chroma vectors. These sequences are expressed according to a fixed time period, and merged by computing and summing their self-similarity matrices. The structural segmentation is obtained by computing and performing a peak-picking on the novelty function from Foote's method [5]. This segmentation is refined by a clustering step relying on a Non-negative Matrix Factorization-based feature space [6].
- System RBH2 by Rocha *et al.* is designed for Electronic Dance Music [7]. After having estimated the first downbeat and the tempo of the piece, the system performs its structural segmentation by a peak-picking of the novelty function from [5] calculated on a beat-synchronized timbral representation of the music signal. A post-processing step adjusts the structural boundaries to favor segments of 8 or 16 bars if the system is confident regarding its tempo estimation.

Systems MHRAF1, MND1, NB2 and WB1 - as well as IRISA11, described in Section II - are mainly based on repetition criteria :

- System MHRAF1 by Martin *et al.* performs several segmentations of a music piece at different time scales [8]. First, the longest repetitions are searched within the sequence of Chroma vectors describing the piece. Then, every iteration of the algorithm locates shorter repetitions, leading to the construction of a tree of repeated segments. The selection of a particular level of the tree leads to the structure estimated by the algorithm. Repetitions are detected using a temporal alignment method which is also robust to transpositions. The system considered here is the version submitted in 2012.
- System MND1 by Mauch *et al.* computes a similarity matrix from a sequence of Chroma vectors representing the music piece. This matrix is filtered and the repetitions are searched by localizing sub-diagonal stripes of high similarity. Repetitions are assumed to begin on a downbeat and last a multiple of four beats. The piece is finally segmented by favoring the repetitions of highest similarity and duration [9].
- System NB2 by Nieto and Bello [10] first performs the structural segmentation of a piece using the approach used in SMGA1 (described later in this section). It is then refined with a K-means clustering where the tonal content of each resulting segment is represented by the 2D Fourier Magnitude Coefficients computed on a variant of its Chroma features.

[1]Except one (OYZS1) which has not been released to our knowledge.

- System WB1 by Weiss and Bello performs the structural segmentation of a music piece by a Shift-Invariant Probabilistic Latent Component Analysis of its beat-synchronous chromagram. First, a learning step builds a dictionary of atoms, namely sequences of Chroma vectors of fixed duration. Second, a Viterbi algorithm is used to find the best sequence of atoms describing the piece's sequence of Chroma vectors [13].

Systems GS1, MP2 and SMGA1 - as well as IRISA10 and IRISA11 in Section II - use a more or less complex combination of several segmentation criteria, or rely on new ones :

- System GS1 implements the supervised-learning approach by Grill and Schlüter mentioned in Section III-B of the article. It implements a convolutional neural network trained on a subset of the SALAMI dataset (see Section V-A of the article) enriched with annotations produced with the same guidelines to "decide [for each short excerpt of the music piece] whether there is a structural boundary at its center or not" [14].
- System MP2 by McFee and Ellis estimates the structural boundaries of the current song using a combination of homogeneity and repetition criteria. A music piece is represented with a sequence of beat-synchronous feature vectors resulting from the concatenation of timbral and tonal features (MFCCs, Chroma vector) along with two sets of "structural features" inspired from the approach used in SMGA1 [12] which encode repetitions of the timbral and tonal content over time, and four beat-related features. The resulting sequence of high-dimensional feature vectors is then adjusted following an adapted Fisher linear discriminant analysis, and segmented using an agglomerative clustering which favors the grouping of temporally close features. The decision for stopping the clustering relies on an AIC-based function [15].
- System SMGA1 by Serrà *et al.* first represents the music piece by its sequence of Chroma vectors. Each Chroma vector is enriched with its immediate temporal predecessor in order to "emulate short-time memory" [11]. This representation is used to compute a recurrence plot turned into a time-lag matrix filtered by a 2D Gaussian kernel. Then, every column of the filtered time-lag matrix is compared with its immediate temporal successor by a Euclidean distance and a peak-picking is performed on the resulting curve to get the segment boundaries. Such an approach somehow mixes homogeneity and repetition criteria [12].

II. Extended description of the three IRISA systems (Section IV-A of the article)

This appendix describes the components of systems IRISA10, IRISA11 and IRISA12, designed and evaluated in the scope of MIREX between 2010 and 2012 under the names BV1, SBVRS1 and SBV1. They all implement a regularity constraint and rely on the Viterbi algorithm presented in Section IV-B.

A. IRISA10

This system is motivated by the observation of structural cues of various nature across the music pieces: timbral homogeneity, harmonic repetition, and "punctuation marks" at the end of segments (local timbral fluctuations such as brief sound effects, drum fills or silence). The idea was therefore to implement a multi-criteria approach using timbral and tonal numeric features under an experimental model of regularity constraint [16].

Features: The timbral and tonal properties of the music piece are respectively described through sequences of MFCCs and Chroma vectors calculated at the beat rate[2]. We consider 20 MFCCs including the 0th coefficient[3] and Chroma vectors of size 12[4].

Data distortion cost: Three segment detection criteria are used to search for segments according to the aforementioned structural cues: a homogeneity breakdown criterion and an event detection criterion calculated on the MFCCs and a repetition breakdown criterion calculated on the Chroma vectors. They are all expressed through a same probabilistic framework, the Generalized Likelihood Ratio (GLR), in order to be simply combined linearly and form the Φ cost in section IV-A.

The GLR compares the likelihood of two antagonistic assumptions, H_0 and H_1, on the probability distribution of a given sequence of observations x:

$$\text{GLR} = \frac{P(x|H_1)}{P(x|H_0)} = \frac{\sup_{\theta\in\Theta_1} p(y|\theta)}{\sup_{\theta\in\Theta_0} p(x|\theta)} \tag{1}$$

where Θ_0 and Θ_1 are two subsets of Θ, the space of parameters of probability distributions. Thus, a large value of the GLR implies that H_1 is plausible.

Our three criteria are computed by means of a log(GLR) on a sliding window centered on the time frame where the breakdown is tested. We choose H_1 as the breakdown assumption and H_0 as the "non-breakdown" assumption, so as to correlate peaks of the criteria to a high probability of occurrence of structural boundaries. Let $x^0 = \{x_t\}_{1\leq t\leq 2N}$ be the sequence of feature vectors representing the temporal neighborhood of a particular time index in a music piece, $N \in \mathbb{N}$.

- The homogeneity breakdown criterion ϕ_H is calculated by taking H_1 as the assumption that the two halves of x, noted $x^+ = \{x_t\}_{1\leq t\leq N}$ and $x^- = \{x_t\}_{N+1\leq t\leq 2N}$, relate to two different Gaussian distributions $G(\mu^+, \Gamma^+)$ and $G(\mu^-, \Gamma^-)$, and H_0 as the assumption that x is tied to a single Gaussian distribution $G(\mu, \Gamma)$. μ^i and Γ^i represent the mean and the covariance of the aforementioned Gaussian distributions, $i = \{-, +\}$. As shown in [19, p. 141–142], ϕ_H can be written as:

$$\phi_H = N(\log(\det(\Gamma) - \frac{\log(\det(\Gamma^+)) + \log(\det(\Gamma^-))}{2}) \tag{2}$$

[2]The beats are estimated using MATLAB scripts by Ellis [17]
[3]MFCCs are extracted using the MA toolbox developed by Pampalk [18]
[4]Chroma vectors are extracted using MATLAB scripts by Ellis [17]

- The event detection criterion ϕ_E is calculated by taking H_1 as the assumption that $\{x_t\}_{N-L+1\leq t\leq N+L}$ and $\{x_t\}_{t\in[\![1,N-L]\!]\cup[\![N+L+1,2N]\!]}$ relate to two different Gaussian distributions, $L < N$, and H_0 as the assumption that x is tied to a single Gaussian distribution. The formula used to compute the log(GLR) is therefore the same than for the homogeneity breakdown criterion with $x^+ = \{x_t\}_{N-L+1\leq t\leq N+L}$ and $x^- = \{x_t\}_{t\in[\![1,N-L]\!]\cup[\![N+L+1,2N]\!]}$.
- The repetition breakdown criterion ϕ_R is calculated with H_0 assuming the repetition of x elsewhere in the current music piece, and H_1 assuming that the sequences $x^+ = \{x_t\}_{1\leq N}$ and $x^- = \{x_t\}_{N+1\leq 2N}$ can't be found contiguously elsewhere within the piece ("no-repetition" assumption). In this case, the maximization of likelihoods implies the obtention of sequences of features y, y^+ and y^- modeling the best x, x^+ and x^- respectively. This research is performed by the calculation of Euclidean distances between sequences of the same size, which is equivalent to consider Gaussian distributions with fixed variances. Thus, ϕ_R can be expressed as follows [19, p. 141–142] :

$$\phi_R \propto -\sum_{t=1}^{N}\|y_t^+ - x_t^+\|^2 - \sum_{t=1}^{N}\|y_t^- - x_t^-\|^2 + \sum_{t=1}^{2N}\|y_t - x_t\|^2 + constant \quad (3)$$

The three criteria are calculated over the music piece, filtered to keep their dominant peaks [20], normalized[5] and summed up to obtain a global breakdown criterion. We calculate the data distortion cost $\Phi_{\text{IRISA10}}(s)$ by summing up all the values taken by the global criterion within temporal segment s. In this way, a high cost is obtained for segments containing dominant peaks from ϕ_H, ϕ_R and ϕ_E.

Constraint: The following structural deviation cost function was our first attempt to model a regularity constraint within the structural segmentation process. We defined it as:

$$\Psi^0(s) = \frac{m}{\tau} + \frac{\tau}{m} - 2 \quad (4)$$

where s is a segment of size m and τ is the structural period.

The minimum is reached when $m = \tau$. Ψ^0 is asymmetric w.r.t. τ in order to apply a softer penalization for segments whose size is below the structural period. τ is estimated by performing a FFT on the filtered homogeneity breakdown criterion, and selecting the frequency of highest energy whose period is close to a target value[6] set to $\sqrt{T}$.

B. IRISA11

This system combines a repetition criterion calculated from chords estimated from the audio, with a new model of regularity constraint [22]. The use of symbolic features was motivated by the will to incorporate information from other MIR tools in the structural segmentation task. We chose an analytical formulation of the regularity cost which allows a broader range of constraint behaviors compared to the one in IRISA10.

Features: Each music piece is described by a sequence of estimated chords[7] expressed according to the scale of downbeats and onbeats[8]. Each chord label is associated to a distinct symbol, so as to perform exact comparisons between the chords in the following.

Data distortion cost: IRISA11 uses a simple data distortion cost Φ_{IRISA11} quantifying the repetitiveness of segments within a music piece. Let $x = \{x_t\}_{1\leq t\leq N}$ be the sequence of symbolic features describing the music piece, and let s be a segment of size m associated to the sequence of features $\{x_t, ..., x_{t+m}\}$. We define:

$$\Phi_{\text{IRISA11}}(s) = \min_{\theta\in Z}\left\{\sum_{p=0}^{m-1} 1 - \delta(x_{t+p}, x_{\theta+p})\right\} \quad (5)$$

where δ is the Kronecker delta : $\delta(x_i, x_j) = 1$ if $x_i = x_j$, and $\delta(x_i, x_j) = 0$ otherwise. We consider the interval $Z = [1, t-m] \cup [t+m, N]$ so as to avoid internal comparisons with segment s.

Constraint: As detailed in Section VI-B, the regularity constraint is modeled using the parametrized cost Ψ_α to study a broader variety of behaviors. It is defined as:

$$\Psi_\alpha(m) = \left|\frac{m}{\tau} - 1\right|^\alpha \quad (6)$$

Ψ_α is non-convex if $0 < \alpha < 1$, and it is convex if $\alpha > 1$. In the settings of MIREX 2011, α was set to 0.5, and λ was tuned using the MIREX10 dataset.

C. IRISA12

This system is composed in the same way as IRISA11 except for the tonal features used and the data distortion cost, which relies on the inner organization of structural segments. Following the work of Bimbot et al. [26], the musical content unfolds over time using a particular logic, e.g. by means of local repetition or alternation of mid-term entities forming patterns like $aabb$ or $abab$. In a number of cases, this logic, which brings the listener to expect how the musical flow will behave, is broken at the end of the segment by the appearance of a new mid-term entity whose content is less predictable. This "explains" frequent patterns such as $aabc$ or $abac$. We therefore designed the data distortion cost of IRISA12 so as to detect the first mid-term entity of structural segments, measuring if it is repeated during its close future and contrasts with the previous entity.

Features: The music piece is represented as a sequence of Chroma vectors[9] expressed at the same scale than for IRISA11.

[5]Each dominant peak is associated its image value according to the complementary error function erfc, which is an empirical way to bring the peaks' amplitudes of different criteria to comparable values.

[6]Denoting T the length of the sequence of feature vectors for the entire music piece, $\sqrt{T}$ minimizes the *predominant informative context* criterion as explained in [21].

[7]The chord estimation is performed by the algorithm by Ueda and al [23], and considers the following chord types beside each possible tonic among 12 pitch classes: major, minor, augmented, diminished, seventh and "no-chord".

[8]In practice, we chose the timescale synchronous to the beat and downbeat scales whose period is close to 1 s. Beats and downbeats were estimated using MATLAB scripts by Davies [24], [25].

[9]This time we used the *Chroma Pitch* features extracted from Chroma Toolbox [27].

Data distortion cost: IRISA12 relies on a segment detection criterion measuring if the neighboring features of each time frame coincide with the first entity of a structural segment. Let $x = \{x_t\}_{1 \leq t \leq 4N}$ be the sequence of Chroma vectors within the analysis window, we divide it into four entities of four feature vectors each and note them $x^i = \{x_t\}_{iN+1 \leq t \leq (i+1)N}$, $i \in \{0, 1, 2, 3\}$. We check whether the first entity of a structural segment coincides with x^1 using the following criterion :

$$\Phi_{\text{IRISA12}} = \lambda_1 \sigma_{\text{Repeat}} + \lambda_2 \sigma_{\text{Contrast}} \tag{7}$$

σ_{Repeat} is a cost measuring how x^1 repeats within x^2 and x^3. Noting $z_j = (x^j - x^1)$ the difference between the j^{th} entity and the first one, $j \in \{2, 3\}$, we propose to define this cost as follows :

$$\sigma_{\text{Repeat}} = \frac{\sum_{l=1}^{N} \min(z_2(l), z_3(l))}{\|x^1\|^2} \tag{8}$$

Looking for the minimal distance between the coefficients of x^j and the assumed first entity x^1 accounts for the inner organization of a structural segment.

σ_{Contrast} is a cost quantifying the difference between x^1 and x^0. Indeed, if x^1 coincides with the first entity of a structural segment, it is probable that its previous element contrasts with it, e.g. as contiguous segments of inner structure $abab$, $abac$,... do. We experimentally chose to compute this cost with a *cotan* function as it performed well in comparison to other functions of similar behavior:

$$\sigma_{\text{Contrast}} = \text{cotan}(x^0, x^1) \tag{9}$$

Finally, these two costs are balanced using tuning parameters λ_1 and λ_2, which take their values in $\mathbb{R}^+$.

Constraint: This system uses the same regularity constraint than IRISA11, i.e., Ψ_α. The parameters were set to $\lambda_1 = 1$, $\lambda_2 = 0.04$, $\lambda = 0.41$ and $\alpha = 0.93$.

REFERENCES

[1] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 318–326, Feb. 2008.

[2] C. Cannam, E. Benetos, M. Mauch, M. E. P. Davies, S. Dixon, C. Landone, K. Noland, and D. Stowell, "MIREX 2015: VAMP Plugins from the Centre for Digital Music," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, October 2015.

[3] R. Chen and M. Li, "Music structural segmentation by combining harmonic and timbral information," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Oct 2011, pp. 477–482.

[4] G. Peeters, "MIREX 2010 music structure segmentation task : IRCAMsummary submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, Oct. 2010.

[5] J. T. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proceedings of the 2000 IEEE International Conference on Multimedia and Expo (ICME)*, Aug. 2000, pp. 452–455.

[6] F. Kaiser, T. Sikora, and G. Peeters, "MIREX 2012 - Music structural segmentation task : IRCAMstructure submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2012.

[7] B. Rocha, N. Bogaards, and A. Honingh, "Segmentation and timbre similarity in electronic dance music," in *Proceedings of the Sound and Music Computing Conference (SMC 2013)*, 2013, pp. 754–761.

[8] B. Martin, P. Hanna, M. Robine, and P. Ferraro, "Structural analysis of harmonic features using string matching techniques (extended abstract)," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2011.

[9] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009, pp. 231–236.

[10] O. Nieto and J. P. Bello, "MIREX 2014 entry: 2D Fourier magnitude coefficients," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2014.

[11] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "Unsupervised detection of music boundaries by time series structure features," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, July 2012, pp. 1613–1619.

[12] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos, "The importance of detecting boundaries in music structure annotation," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, oct 2012.

[13] R. Weiss and J. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proceedings of the 11th International Society on Music Information Retrieval (ISMIR)*, October 2010, pp. 123–128.

[14] T. Grill and J. Schlüter, "Structural segmentation with convolutional neural networks MIREX submission," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, October 2015, p. 3 pages.

[15] B. McFee and D. P. Ellis, "DP1, MP1, MP2 entries for MIREX 2013 structural segmentation and beat tracking," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, 2013.

[16] G. Sargent, F. Bimbot, and E. Vincent, "A structural segmentation of songs using generalized likelihood ratio under regularity assumptions (extended abstract)," in *Proceedings of the Music Information Retrieval Evaluation eXchange (MIREX)*, october 2010.

[17] D. P. Ellis, "Music beat tracking and cover song identification (last accessed: Sept. 2016)," 2007. [Online]. Available: http://labrosa.ee.columbia.edu/projects/coversongs/

[18] E. Pampalk, "Ma toolbox (last accessed: Sept. 2016)," 2007. [Online]. Available: http://www.pampalk.at/ma/

[19] G. Sargent, "Estimation de la structure de morceaux de musique par analyse multi-critères et contrainte de régularité," Ph.D. dissertation, Université de Rennes 1, 2013.

[20] M. Seck, R. Blouet, and F. Bimbot, "The IRISA/ELISA Speaker Detection and Tracking Systems for the NIST'99 Evaluation Campaign," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 154–171, Jan. 2000.

[21] F. Bimbot, O. Le Blouch, G. Sargent, and E. Vincent, "Decomposition into autonomous and comparable blocks: A structural description of music pieces," in *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, Aug. 2010, pp. 189–194.

[22] G. Sargent, F. Bimbot, and E. Vincent, "A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Oct. 2011.

[23] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, "HMM-based approach for automatic chord detection using refined acoustic features," in *Proceedings of the 2010 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010, pp. 5506–5509.

[24] M. E. P. Davies, "Towards automatic rhythmic accompaniment," Ph.D. dissertation, Department of Electronic Engineering, Queen Mary, University of London, 2007.

[25] A. M. Stark, M. E. P. Davies, and M. D. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx)*, September 2009.

[26] F. Bimbot, E. Deruty, G. Sargent, and E. Vincent, "System & contrast : A polymorphous model of the inner organization of structural segments within music pieces," *Music Perception*, vol. 32, no. 5, pp. 631–661, 2016, this paper extends and deepens former work described in IRISA PI-1999 [Research Report], December 2012, 40 pages, hal-00868398.

[27] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, October 2011, pp. 215–220.